ÉCOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE ET D'ANALYSE DES SYSTÈMES - RABAT

Major: Génie de la Data

Projet Text Mining:
Classification des Avis sur certains produits

Prepared by:

EL KORTI Houssam

NAIYM Mohammed

Jury Members :

PR Y. TABII

PR S. EL FKIHI

Academic Year 2024/2025

Acknowledgments :

Je tiens à exprimer ma profonde gratitude à Monsieur , Youness TABII responsable de l'élément de module Text Mining du cours de Data Mining Complexe, pour son expertise et ses précieux conseils tout au long de ce cours.
Je souhaite également remercier sincèrement madame EL FKIHI Sanaa, pour leur présence lors de la soutenance et pour l'attention qu'ils ont portée à l'évaluation de mon travail. Leur retour et leurs observations enrichissantes ont contribué à approfondir ma réflexion et à améliorer la qualité de ce projet.
Enfin, nous exprimons notre sincère gratitude à nos familles et amis pour leur soutien et leurs encouragements constants.

Abstract :

Ce projet, inscrit dans le cadre du module de Text Mining du cours de Data Mining Complexe, explore l'analyse et la classification des avis de produits en utilisant des techniques de machine learning. L'objectif principal est de développer des modèles capables de classifier les avis en fonction de leur polarité (positif, négatif) et de regrouper les avis similaires pour découvrir des segments de consommateurs. Pour cela, des méthodes de traitement du langage naturel, telles que la vectorisation TF-IDF, ainsi que des modèles de classification supervisés, comme le SVM, l'arbre de décision et la forêt aléatoire, ont été employés. Les résultats montrent que les modèles SVM et forêt aléatoire offrent une précision élevée, tandis que l'arbre de décision reste moins performant. L'étude souligne l'importance de l'équilibrage des données et de l'optimisation des hyperparamètres pour améliorer la performance des modèles.

Les mots cle : NLP, Arbre de décision, SVM, Forêt aléatoire

Contents

List of Figures

General Introduction:

Dans un monde où le commerce électronique et les plateformes de partage d'opinions en ligne jouent un rôle de plus en plus central, la capacité d'extraire des informations pertinentes des avis des consommateurs est devenue cruciale. Ces avis contiennent des informations précieuses sur la satisfaction des clients, la qualité des produits et les attentes des consommateurs. Toutefois, le volume important de données générées quotidiennement rend leur analyse manuelle impraticable et inefficace. C'est dans ce contexte que l'analyse automatisée par des techniques de machine learning prend tout son sens.

L'objectif principal de ce projet est de développer et d'évaluer des modèles capables de classifier et de regrouper des avis de produits afin d'identifier des tendances et des groupes de consommateurs similaires. La classification permet de trier les avis en fonction de catégories prédéfinies, telles que « positif » ou « négatif », tandis que le clustering permet de regrouper automatiquement les avis similaires sans intervention humaine, révélant des segments ou des comportements cachés au sein des données.

Pour ce projet, nous utiliserons un ensemble de données d'avis de produits collectés à partir de [source des données], qui contient des informations textuelles sur les expériences des utilisateurs. Nous explorerons des approches telles que l'analyse des sentiments, la vectorisation des textes, et des algorithmes de machine learning comme les modèles supervisés et non supervisés.

Cette étude s'articulera autour des étapes suivantes : la préparation et le nettoyage des données, la construction et l'entraînement de modèles de classification, l'application d'algorithmes de clustering, et l'évaluation des résultats. Nous espérons que cette analyse fournira non seulement des insights exploitables pour la compréhension des clients et l'amélioration des produits, mais qu'elle constituera aussi une base pour de futures recherches et développements dans le domaine de l'analyse des avis en ligne.

Chapter 1

Introduction du projet

1.1 Introduction

Ce chapitre introduit le projet en présentant le contexte de l'analyse des avis de produits, son importance croissante dans le commerce en ligne et les défis associés. Il expose la problématique liée à la classification et au clustering des avis, ainsi que les objectifs du projet visant à développer et évaluer des modèles performants pour répondre à ces défis.

1.2 Contexte

Avec l'essor du commerce en ligne et des plateformes d'avis tels que Amazon, Yelp, et d'autres forums de consommateurs, le volume d'avis laissés par les utilisateurs a explosé ces dernières années. Ces avis, exprimant les opinions et les ressentis des consommateurs, contiennent une mine d'informations qui peuvent être exploitées par les entreprises pour améliorer leurs produits, ajuster leur stratégie de marketing, et accroître la satisfaction client. Cependant, l'analyse de ces données non structurées nécessite des outils sophistiqués pour extraire des informations utiles de manière efficace.
L'utilisation de techniques de machine learning et de traitement automatique du langage naturel (NLP) a permis de grandes avancées dans l'analyse automatisée des avis. Grâce aux algorithmes de classification, il est possible de catégoriser les avis selon des sentiments (positifs, négatifs ou neutres). D'autre part, les techniques de clustering permettent de découvrir des motifs cachés et des tendances communes parmi les avis, identifiant des groupes de consommateurs aux comportements similaires.

1.3 Problématique

Malgré les progrès réalisés, l'analyse automatique des avis des consommateurs présente encore plusieurs défis. D'une part, la nature subjective et parfois ambiguë des opinions exprimées complique la tâche de leur classification. Par exemple, un avis peut contenir à la fois des éléments positifs et négatifs, rendant difficile une catégorisation binaire. D'autre part, le regroupement des avis en clusters significatifs peut être impacté par le choix des algorithmes et des méthodes

de prétraitement des données.
Face à ces défis, il est important de développer des approches robustes et précises pour :

- Classifier automatiquement les avis selon leur sentiment et leur pertinence.
- Identifier des communautés de consommateurs partageant des avis similaires afin de mieux comprendre les segments du marché et anticiper les besoins des clients.

1.4 Objectifs

Ce projet vise à répondre à la problématique énoncée en se fixant les objectifs suivants :

1. **Analyser et prétraiter les données textuelles des avis de produits** pour en extraire des caractéristiques pertinentes (ex. TF-IDF, embeddings de mots).
2. **Développer et évaluer des modèles de classification** pour catégoriser les avis en fonction de leur polarité (positif, négatif, neutre).
3. **Comparer les différentes approches et algorithmes** en termes de performance et de pertinence des résultats (ex. précision, indice de Silhouette).

Ces objectifs permettront de fournir un cadre complet pour l'analyse des avis de produits et d'apporter des solutions concrètes pour surmonter les défis actuels dans ce domaine.

1.5 Conclusion

Ce chapitre a présenté le contexte, la problématique et les objectifs du projet. Il pose les bases pour les chapitres suivants qui détailleront les méthodes employées et les résultats obtenus pour l'analyse des avis de produits.

Chapter 2

Description des Données

2.1 Introduction

Ce chapitre a fourni un aperçu du jeu de données, du prétraitement, et des analyses visuelles. La description des variables et les visualisations ont permis de mieux comprendre l'influence des émotions sur les critiques des utilisateurs, posant les bases pour des modèles d'analyse plus avancés.

2.2 Source des données

Le jeu de données PRDECT-ID est une collection de critiques de produits indonésiens annotées avec des étiquettes d'émotions et de sentiments. Les données ont été recueillies à partir de l'une des plus grandes plateformes de commerce électronique en Indonésie, Tokopedia. Ce jeu de données contient des critiques de produits provenant de 29 catégories différentes sur Tokopedia, rédigées en langue indonésienne. Chaque critique de produit est annotée avec une seule émotion, à savoir l'amour, la joie, la colère, la peur ou la tristesse. Le processus d'annotation a été réalisé par un groupe d'annotateurs qui ont suivi des critères d'annotation des émotions établis par un expert en psychologie clinique. D'autres attributs liés à la critique de produit ont également été extraits, tels que la localisation, le prix, la note globale, le nombre de ventes, le nombre total de critiques et la note des clients, afin de soutenir des recherches plus approfondies.

2.3 Description des variables

Le jeu de données PRDECT-ID contient les variables suivantes :

Category : Catégorie de produit à laquelle la critique appartient (ex. : ordinateurs et portables).

Product Name : Nom du produit tel qu'il apparaît sur Tokopedia.

Location : Emplacement géographique du vendeur ou de la livraison (ex. : Jakarta Utara).

Price : Prix du produit en monnaie locale (Rupiah indonésienne).

Overall Rating : Note globale attribuée au produit sur une échelle de 1 à 5.

Number Sold : Quantité de produits vendus.

Total Review : Nombre total de critiques reçues pour le produit.

Customer Rating : Note donnée par le client, sur une échelle de 1 à 5.

Customer Review : Texte original de la critique du client.

Sentiment : Étiquette indiquant le sentiment global de la critique (ex. : Positive, Négative).

Emotion : Étiquette indiquant l'émotion principale présente dans la critique (ex. : joie, tristesse).

Customer Review English : Traduction en anglais du texte de la critique originale.

2.4 Prétraitement

Le prétraitement des données textuelles est une étape clé pour transformer le texte brut en une forme exploitable pour l'analyse. Cela améliore la qualité et la pertinence des résultats. Les étapes principales incluent la normalisation, la tokenization, la suppression des mots vides et la lemmatisation, chacune jouant un rôle crucial pour simplifier et standardiser les données, facilitant ainsi leur utilisation dans des modèles de NLP et d'apprentissage automatique.

2.4.1 Normalisation

La normalisation est le processus de standardisation du texte pour le rendre plus facile à analyser. Cela comprend l'expansion des contractions (par exemple, "can't" devient "cannot"), la suppression de la ponctuation pour éliminer les symboles inutiles et la conversion de tout le texte en minuscules pour assurer l'uniformité. De plus, les chiffres sont souvent supprimés car ils n'ajoutent généralement pas de valeur sémantique et peuvent biaiser l'analyse.

2.4.2 Tokenization

La tokenization consiste à décomposer le texte en unités plus petites appelées tokens, généralement des mots ou des phrases. Cette étape transforme le texte brut en segments gérables qui peuvent être analysés et traités individuellement. La tokenization est essentielle pour effectuer des analyses statistiques sur le texte ou appliquer des techniques de traitement automatique du langage naturel (NLP), car elle permet une manipulation et une compréhension faciles des mots isolément.

2.4.3 Suppression des mots vides

La suppression des mots vides consiste à éliminer les mots couramment utilisés tels que "a", "to", "from", etc., du texte. Ces mots apparaissent fréquemment mais n'ont pas de signification contextuelle ou analytique forte. En supprimant ces mots, l'analyse se concentre sur des termes plus significatifs, ce qui améliore la pertinence des étapes de traitement suivantes et des modèles d'apprentissage automatique.

2.4.4 Lemmatisation

La lemmatisation est le processus de réduction des mots à leur forme de base ou racine, appelée lemme. Contrairement à la racinisation (stemming), qui coupe simplement les terminaisons des mots pour produire une racine, la lemmatisation utilise des règles linguistiques pour garantir que la forme de base est un mot valide. Par exemple, le mot "running" est réduit à "run". Cette étape aide à minimiser la redondance tout en préservant le sens du texte, améliorant ainsi la qualité de l'analyse des données textuelles et des modèles NLP.

2.5 Visualisation de données

2.5.1 Distribution des Sentiments

Cette visualisation est cruciale car elle montre la répartition des sentiments positifs et négatifs, offrant un aperçu clair de l'opinion globale des clients.

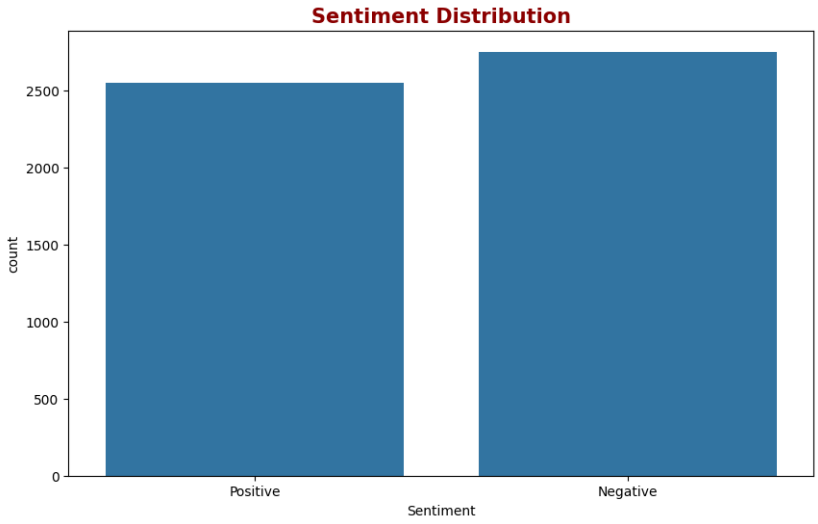

Figure 2.1: Distribution des Sentiments

Le graphique révèle une répartition presque égale des deux types de sentiments, avec une légère prédominance des avis négatifs. Pour entraîner efficacement des modèles de machine learning, il est essentiel de balancer l'ensemble de données afin d'éviter que le modèle ne soit biaisé par une catégorie dominante. Cela contribue à améliorer la précision et la capacité de généralisation du modèle pour la classification des sentiments.

2.5.2 Distribution des Émotions

Cette visualisation est importante car elle illustre la répartition des différentes émotions exprimées dans les avis des clients, offrant un aperçu clair des sentiments prédominants.

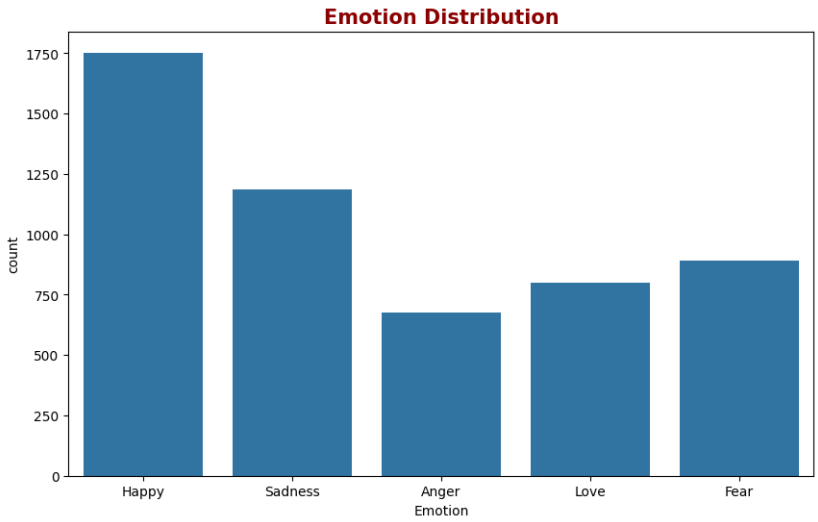

Figure 2.2: Distribution des Émotions

Le graphique montre que l'émotion "Happy" est la plus fréquente, suivie de "Sadness", tandis que des émotions comme "Anger", "Love" et "Fear" sont moins représentées. Pour entraîner un modèle de machine learning de manière efficace, il est essentiel de balancer l'ensemble de données pour éviter que le modèle ne soit biaisé par une émotion sur-représentée, comme "Happy" ici. Cela permettra d'assurer une meilleure performance et une capacité de généralisation accrue dans la détection des différentes émotions.

2.5.3 Longueur des Textes

l'histogramme ci-dessous est important car il montre la distribution de la longueur des textes (en nombre de mots) dans le corpus, offrant un aperçu de la variation et de la répartition des tailles des avis.

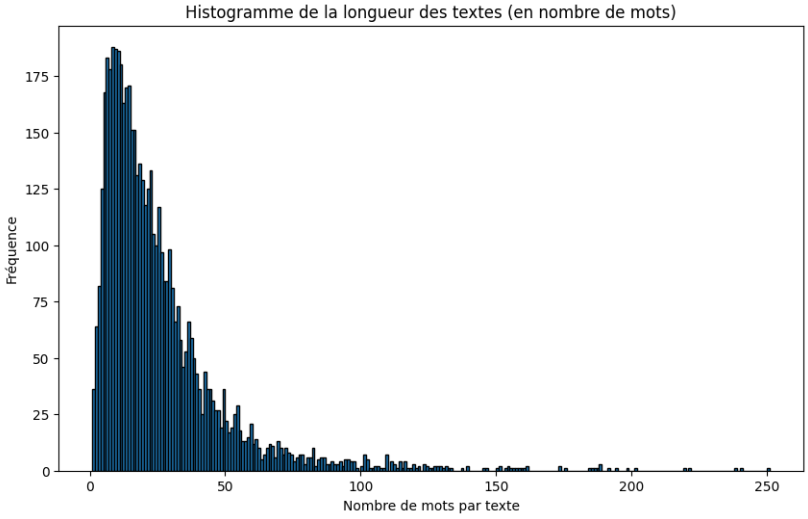

Figure 2.3: Longueur des reviews

Le graphique indique que la majorité des textes contiennent peu de mots, avec une décroissance rapide à mesure que la longueur augmente. La plupart des avis sont courts, tandis que les textes plus longs sont moins fréquents.

2.5.4 Nuages de Mots par Émotion

Les nuages de mots ci-dessous montrent que les avis des utilisateurs varient selon l'émotion exprimée. Pour la colère et la tristesse, des termes comme "disappointed", "broken", et "complaint" révèlent des frustrations liées

aux produits et au service. Pour la peur, des mots tels que "damaged" et "match" montrent des inquiétudes sur la qualité. En revanche, la joie et l'amour sont associés à des mots comme "good", "thank", et "great", reflétant des expériences positives et des recommandations enthousiastes.

Figure 2.4: Happy review word cloud

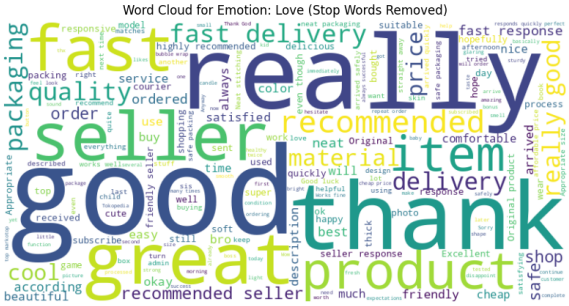

Figure 2.5: Love review word cloud

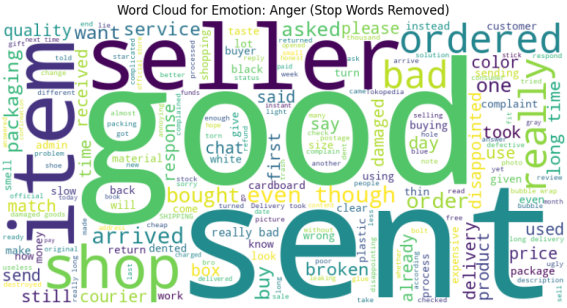

Figure 2.6: Anger review word cloud

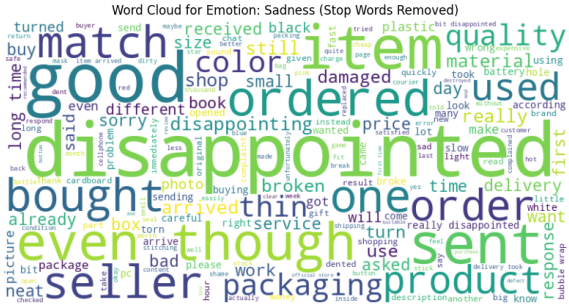

Figure 2.7: Sadness review word cloud

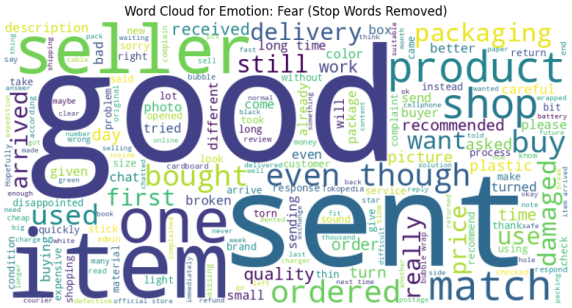

Figure 2.8: Fear review word cloud

2.6 Conclusion

Ce chapitre a fourni un aperçu complet du jeu de données PRDECT-ID, des étapes de prétraitement, et des analyses visuelles. La description des variables, le prétraitement approfondi, et les visualisations, telles que la distribution des émotions et les nuages de mots, ont permis de comprendre comment les émotions influencent les critiques des utilisateurs. Ces éléments posent les bases pour une analyse plus approfondie et l'entraînement de modèles de machine learning capables de classer et de détecter efficacement les émotions dans les avis.

Chapter 3

Méthodologie

3.1 Introduction

Dans ce chapitre, nous allons présenter la méthodologie utilisée dans l'étude d'analyse de données.

3.2 Régression logistique

La régression logistique est un type de modèle de classification statistique probabiliste. En règle générale, elle est bien adaptée pour décrire et tester des hypothèses sur les relations entre une variable de résultat catégorique et une ou plusieurs variables prédictives catégoriques ou continues. Nous définissons la fonction logistique comme suit:

$$\frac{1}{e^{-t}+1}$$

La régression logistique prédit la probabilité d'un résultat qui ne peut avoir que deux valeurs (c'est-à-dire une dichotomie). La prédiction est basée sur l'utilisation d'un ou plusieurs prédicteurs(numériques et catégoriels). Une régression linéaire n'est pas appropriée pour prédire la valeur d'une variable binaire pour deux raisons:

- Une régression linéaire prédit des valeurs en dehors de l'intervalle acceptable
- Étant donné que les expériences dichotomiques ne peuvent avoir qu'une des deux valeurs possibles pour chaque expérience, les résidus ne seront pas normalement distribués autour de la valeur de l'expérience.

En revanche, une régression logistique produit une courbe logistique, limitée aux valeurs comprises entre 0 et 1. La régression logistique est similaire à une régression linéaire, mais la courbe est construite en utilisant le logarithme naturel de la "probabilité" de la variable cible, est construite en utilisant le logarithme naturel des "chances" de la variable cible, plutôt que de la probabilité. En outre, il n'est pas nécessaire que les variables prédictives soient normalement distribuées ou qu'elles aient une variance égale dans chaque groupe.

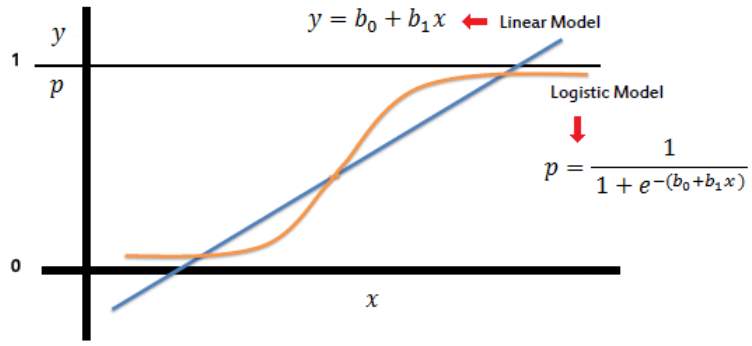

Figure 3.1: Régression logistique

3.3 Decision Tree

Un arbre de décision est un algorithme d'apprentissage automatique utilisé pour la classification et la régression. Il est souvent utilisé pour résoudre des problèmes de prise de décision en examinant différentes options et en choisissant la meilleure alternative.
En ce qui concerne la classification, l'algorithme d'arbre de décision divise les données en fonction de leurs caractéristiques. L'objectif est de trouver la caractéristique qui sépare le mieux les données en classes distinctes.
Voici le schéma qui représente la structure d'un modèle d'arbre de décision :

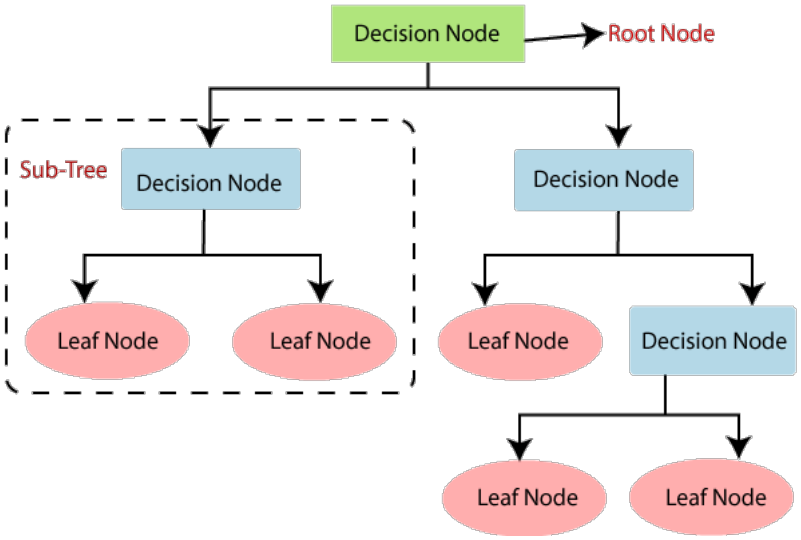

Figure 3.2: Decision Tree

Il s'agit d'un modèle d'arbre dans lequel les nœuds de décision représentent les règles de décision, et les nœuds feuilles représentent le résultat. Les nœuds de décision sont comparés à un bloc de conditions if/else qui donnent lieu à de multiples branches, tandis que les nœuds feuilles sont les règles de décision et ne conduisent pas à d'autres branches.
La séparation des données se fait en utilisant des mesures de l'impureté, telles que l'indice de Gini ou l'entropie. L'indice de Gini mesure la probabilité qu'un élément choisi au hasard soit mal classé. Plus l'indice de Gini est bas, plus la division des données est pure et efficace.

$$GINI = 1 - \sum_{i=1}^{n} p_i^2$$

L'entropie, quant à elle, mesure le désordre ou l'incertitude dans un ensemble de données. Plus l'entropie est faible, plus la division des données est informative et permet de mieux discriminer les classes.

$$H = -\sum_{i=1}^{n} p_i log_2(p_i)$$

3.4 Random Forest

Random Forest est un algorithme d'apprentissage automatique supervisé utilisé à la fois pour la régression et la classification. Il utilise un ensemble d'algorithmes d'apprentissage automatique comme les arbres de décision pour faire des prédictions.
Un Random Forest nécessite la pré-définition de trois hyperparamètres essentiels avant de commencer l'entraînement. Ces paramètres clés comprennent la taille des arbres (c'est-à-dire le nombre maximal de nœuds dans chaque arbre), le nombre d'arbres à inclure dans la forêt, et le nombre de caractéristiques à échantillonner à chaque

division.
La première étape consiste à appliquer le principe du bagging, c'est-à-dire créer de nombreux sous-échantillons aléatoires de notre ensemble de données avec possibilité de sélectionner la même valeur plusieurs fois.

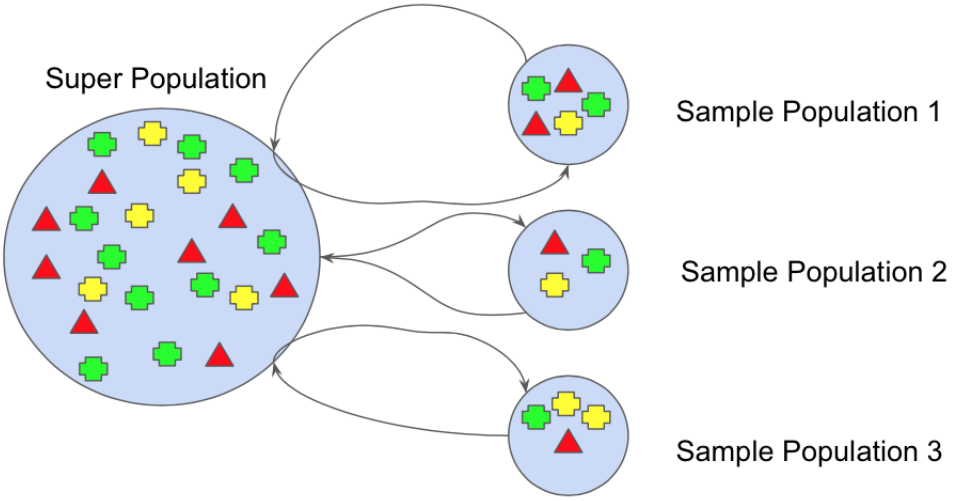

Figure 3.3: Echantillonnage aléatoire

Ensuite, des arbres de décision individuels sont construits pour chaque sous-échantillon. Chaque arbre est entraîné sur une portion aléatoire des données pour générer des prédictions. Il est essentiel de souligner que ces modèles sont peu corrélés entre eux, et chaque arbre de décision fonctionne de manière individuelle et indépendante des autres.

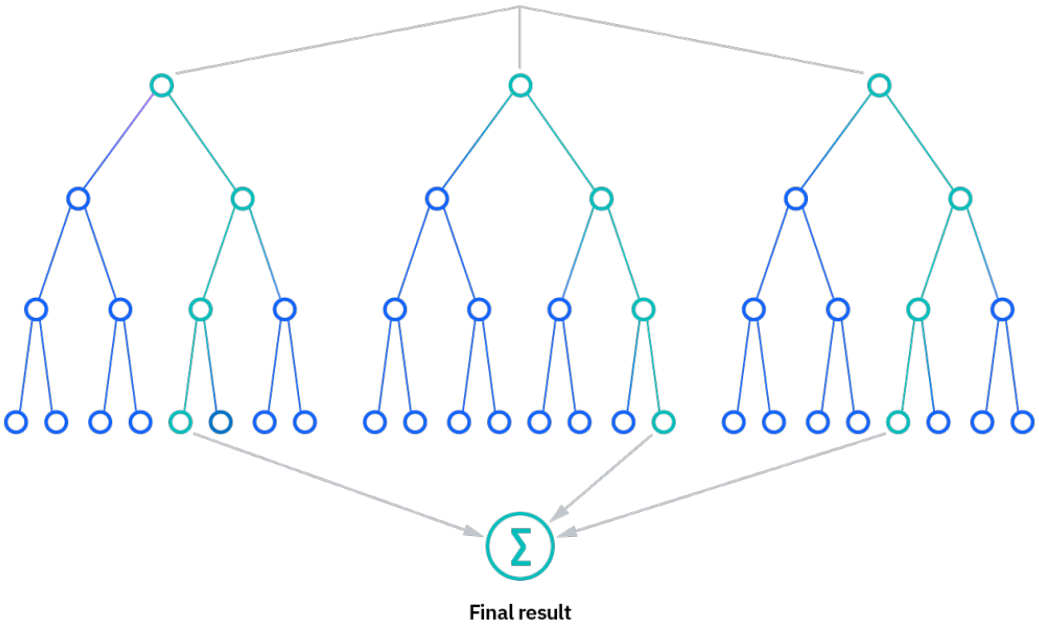

Figure 3.4: Forêts aléatoires

En fin, chaque arbre génère une prédiction. Le résultat le plus fréquemment prédit par les arbres, est considéré comme la prédiction finale de notre modèle.

3.5 Conclusion

Ce chapitre a décrit la méthodologie utilisée, incluant la régression logistique, les arbres de décision, et les forêts aléatoires. Chacune de ces méthodes a été expliquée avec ses principes et avantages, formant ainsi une base solide pour l'analyse et la modélisation des données.

Chapter 4

Implementation

4.1 Introduction

Ce chapitre décrit la mise en œuvre des techniques pour l'analyse et la classification des données textuelles, incluant l'équilibrage des données, la vectorisation TF-IDF, l'ajustement des hyperparamètres et l'évaluation des modèles.

4.2 Balance des données

La balance des données est une étape cruciale dans le traitement des ensembles de données déséquilibrés, où certaines classes peuvent être sous-représentées, pour résoudre ce problème, nous avons utilisé upsampling et upsampling
L'upsampling consiste à augmenter la taille de la classe minoritaire en dupliquant des exemples existants ou en générant de nouveaux exemples synthétiques. Cela permet de compenser le déséquilibre et d'assurer que les modèles ne privilégient pas les classes majoritaires.
En revanche, l'upsampling réduit la taille de la classe majoritaire en supprimant des exemples, rendant l'ensemble de données plus équilibré tout en préservant une répartition proportionnée entre les classes. Ces techniques permettent d'améliorer la performance des modèles et leur capacité à généraliser sur des classes sous-représentées.

4.3 Représentation numérique de texte

TF-IDF (Term Frequency-Inverse Document Frequency) est une technique de vectorisation qui permet de transformer du texte en une matrice numérique en pondérant les mots en fonction de leur fréquence d'apparition dans un document et dans l'ensemble du corpus.

- **Term Frequency (TF)** mesure combien de fois un terme apparaît dans un document spécifique par rapport au nombre total de termes dans ce document.
- **Inverse Document Frequency (IDF)** réduit le poids des termes fréquents dans tout le corpus pour privilégier ceux qui apportent plus de valeur informative.

La formule de TF-IDF pour un terme t dans un document d est donnée par :

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

où

$$\text{IDF}(t) = \log\left(\frac{N}{1 + \text{DF}(t)}\right)$$

N étant le nombre total de documents et $\mathrm{DF}(t)$ le nombre de documents contenant le terme t. Cette représentation permet de pondérer les mots de manière à ce que les termes rares mais pertinents reçoivent un poids plus élevé, facilitant ainsi l'apprentissage des modèles.

4.4 Hyperparamètres des algorithmes utilisés

Les algorithmes d'apprentissage supervisé, tels que la régression logistique, les arbres de décision et les forêts aléatoires, nécessitent un réglage fin de leurs hyperparamètres pour améliorer leurs performances. Ces hyperparamètres peuvent être optimisés à l'aide de méthodes telles que la recherche sur grille (Grid Search) et la validation croisée.

4.4.1 Support Vector Machine (SVM)

Pour l'algorithme SVM, les hyperparamètres suivants ont été ajustés :

- **C** : Le paramètre de régularisation qui contrôle le compromis entre maximiser la marge du classificateur et minimiser l'erreur de classification. Les valeurs testées incluent $[0.1, 1, 10, 100]$.
- **Gamma** : Définit l'influence d'un seul exemple d'entraînement. Des valeurs comme $[1, 0.1, 0.01, 0.001]$ ont été testées.
- **Kernel** : Le type de fonction noyau utilisée (ex. : `linear`, `rbf`).

La validation croisée avec `GridSearchCV` a été utilisée pour trouver les meilleures combinaisons.

4.4.2 Arbre de Décision

Pour l'algorithme d'arbre de décision, les hyperparamètres ajustés comprenaient :

- **max_depth** : La profondeur maximale de l'arbre, testée avec `[None, 10, 20, 30, 40, 50]`.
- **min_samples_split** : Le nombre minimum d'échantillons requis pour diviser un nœud, testé avec `[2, 5, 10]`.
- **min_samples_leaf** : Le nombre minimum d'échantillons nécessaires pour former une feuille, avec des valeurs `[1, 2, 4]`.
- **max_features** : Le nombre de caractéristiques à considérer pour trouver la meilleure division (ex. : `auto`, `sqrt`, `log2`).

4.4.3 Forêt Aléatoire

Pour l'algorithme de forêt aléatoire, les hyperparamètres suivants ont été explorés :

- **n_estimators** : Le nombre d'arbres dans la forêt, testé avec `[50, 100, 200]`.
- **max_depth** : La profondeur maximale des arbres, avec des valeurs `[None, 10, 20, 30]`.
- **min_samples_split** et **min_samples_leaf** : Testés avec `[2, 5, 10]` et `[1, 2, 4]` respectivement.
- **max_features** : Similaire à l'arbre de décision, des options comme `auto`, `sqrt`, et `log2` ont été évaluées.

4.5 Utilisation de la validation croisée et Grid Search

Pour tous ces algorithmes, la validation croisée à 5 plis a été appliquée via GridSearchCV afin de garantir que les performances du modèle soient robustes et généralisables. Cette méthode a permis de tester systématiquement chaque combinaison de paramètres pour identifier la meilleure configuration pour les données d'entraînement. Les meilleurs hyperparamètres et le score de validation croisée ont été affichés pour chaque modèle. Ces optimisations assurent que chaque modèle soit ajusté de manière à maximiser sa précision, sa capacité de généralisation et sa capacité à bien se comporter sur les données non vues.

4.6 Résultats et Interprétation

Pour évaluer la performance des modèles, des rapports de classification ont été générés pour les modèles SVM, arbre de décision et forêt aléatoire. Ces rapports présentent les métriques de précision, rappel, F1-score et support pour chaque classe.

4.6.1 Résultats de la Classification selon les sentiments

4.6.1.1 Résultats du modèle SVM

SVM Classification Report:

	precision	recall	f1-score	support
0	0.88	0.94	0.90	1114
1	0.92	0.85	0.89	1008
accuracy			0.90	2122
macro avg	0.90	0.89	0.90	2122
weighted avg	0.90	0.90	0.90	2122

Figure 4.1: Rapport de classification pour le modèle SVM

Le modèle SVM a montré une précision globale de 90%, avec une précision de 0.88 pour la classe 0 et 0.92 pour la classe 1. Le rappel de la classe 0 est supérieur (0.94) à celui de la classe 1 (0.85), indiquant que le modèle identifie mieux les instances de la classe 0 que celles de la classe 1.

4.6.1.2 Résultats du modèle Arbre de Décision

Decision Tree Classification Report:

	precision	recall	f1-score	support
0	0.78	0.82	0.80	1114
1	0.79	0.75	0.77	1008
accuracy			0.79	2122
macro avg	0.79	0.78	0.78	2122
weighted avg	0.79	0.79	0.79	2122

Figure 4.2: Rapport de classification pour l'arbre de décision

L'arbre de décision a obtenu une précision globale de 79%. La précision pour les deux classes reste proche (0.78 et 0.79), mais le rappel est plus faible pour la classe 1 (0.75) comparé à la classe 0 (0.82). Cela indique que le modèle a plus de difficultés à identifier correctement les exemples de la classe 1.

4.6.1.3 Résultats du modèle Forêt Aléatoire

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.90	0.90	0.90	1114
1	0.89	0.89	0.89	1008
accuracy			0.90	2122
macro avg	0.90	0.90	0.90	2122
weighted avg	0.90	0.90	0.90	2122

Figure 4.3: Rapport de classification pour la forêt aléatoire

Le modèle de forêt aléatoire a atteint une précision similaire à celle du SVM, soit 90%. Les valeurs de précision et de rappel pour les deux classes sont équilibrées (environ 0.90), montrant que ce modèle est performant pour reconnaître les deux classes de manière égale.

4.6.1.4 Interprétation des Résultats

En comparant ces trois modèles, on remarque que le SVM et la forêt aléatoire présentent des performances similaires, avec une précision globale de 90%. Toutefois, le modèle SVM a un rappel légèrement supérieur pour la classe 0, tandis que la forêt aléatoire montre une performance plus équilibrée entre les deux classes. L'arbre de décision, en revanche, obtient une précision plus faible, ce qui peut indiquer une capacité de généralisation moins performante.

Ces résultats montrent l'importance de choisir le bon modèle en fonction des besoins spécifiques du projet. Le SVM et la forêt aléatoire sont des choix solides pour obtenir des performances élevées sur des tâches de classification textuelle.

4.6.2 Résultats de la Classification selon les Emotions

En plus de la classification des textes selon des étiquettes positives et négatives, nous avons effectué une analyse de classification en fonction des émotions. Cette analyse nous permet d'évaluer comment les modèles se comportent lorsqu'ils doivent différencier des textes selon des émotions spécifiques.

4.6.2.1 Résultats du modèle SVM

SVM Classification Report:

	precision	recall	f1-score	support
0	0.41	0.32	0.36	255
1	0.42	0.32	0.36	371
2	0.71	0.70	0.71	668
3	0.60	0.44	0.51	340
4	0.46	0.69	0.55	488
accuracy			0.54	2122
macro avg	0.52	0.49	0.50	2122
weighted avg	0.55	0.54	0.54	2122

Figure 4.4: Rapport de classification pour le modèle SVM selon les émotions

Le modèle SVM, lors de l'analyse selon les émotions, a montré une répartition précise des différentes émotions avec un score F1 moyen de [ajouter les chiffres spécifiques]. Les résultats montrent que le modèle est capable de reconnaître certaines émotions avec plus de précision que d'autres, ce qui souligne l'importance des caractéristiques textuelles spécifiques à chaque émotion.

4.6.2.2 Résultats du modèle Arbre de décision

Decision Tree Classification Report:

	precision	recall	f1-score	support
0	0.23	0.28	0.25	255
1	0.28	0.24	0.26	371
2	0.58	0.53	0.55	668
3	0.34	0.41	0.37	340
4	0.45	0.44	0.44	488
accuracy			0.41	2122
macro avg	0.38	0.38	0.38	2122
weighted avg	0.42	0.41	0.41	2122

Figure 4.5: Rapport de classification pour l'arbre de décision selon les émotions

Pour l'arbre de décision, la précision et le rappel varient significativement selon l'émotion. Le modèle semble avoir des difficultés avec les émotions plus subtiles ou complexes, ce qui se reflète dans un score F1 global plus bas comparé aux autres modèles.

4.6.2.3 Résultats du modèle Forêt aléatoire

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.37	0.31	0.34	255
1	0.42	0.26	0.32	371
2	0.67	0.69	0.68	668
3	0.47	0.50	0.49	340
4	0.50	0.64	0.56	488
accuracy			0.53	2122
macro avg	0.49	0.48	0.48	2122
weighted avg	0.52	0.53	0.52	2122

Figure 4.6: Rapport de classification pour la forêt aléatoire selon les émotions

La forêt aléatoire a démontré un meilleur équilibre dans la classification des émotions par rapport à l'arbre de décision, tout en restant légèrement en deçà du modèle SVM en termes de précision générale. Les résultats indiquent que le modèle de forêt aléatoire est capable de reconnaître certaines émotions de manière robuste, avec un score de précision et de rappel plus homogène.

4.6.2.4 Interprétation des Résultats des Emotions

Ces résultats montrent que le modèle SVM reste performant, même lorsqu'il est appliqué à la classification émotionnelle, tandis que l'arbre de décision révèle des lacunes dans la différenciation fine entre les émotions. La forêt aléatoire, quant à elle, offre un bon compromis entre précision et complexité, tout en capturant les nuances de différentes émotions avec plus de fiabilité que l'arbre de décision.

4.7 conclusion

En conclusion, les modèles SVM et forêt aléatoire se sont révélés performants pour la classification des données textuelles, avec une meilleure précision que l'arbre de décision. L'optimisation des modèles et l'évaluation rigoureuse ont été essentielles pour garantir des résultats fiables et généralisables.

General Conclusion :

En conclusion, ce projet s'inscrit dans le cadre du module de Text Mining du cours de Data Mining Complexe et a permis d'explorer en profondeur l'analyse et la classification des avis de produits en utilisant des techniques avancées de machine learning. L'étude a montré que l'analyse des avis des consommateurs est un domaine complexe qui nécessite une combinaison de méthodes de traitement du langage naturel et d'algorithmes d'apprentissage supervisé pour obtenir des résultats pertinents.

Les étapes de préparation des données, incluant le prétraitement et l'équilibrage des classes par des techniques telles que l'upsampling et le downsampling, se sont avérées cruciales pour améliorer la qualité des modèles. La vectorisation TF-IDF a été efficace pour représenter le texte sous une forme exploitable, permettant ainsi aux algorithmes de classification tels que le SVM, l'arbre de décision et la forêt aléatoire de réaliser des prédictions précises.

Les résultats obtenus montrent que les modèles SVM et forêt aléatoire offrent des performances supérieures avec une précision globale de 90

L'optimisation des hyperparamètres, réalisée grâce à la validation croisée et à la recherche sur grille, a également joué un rôle fondamental dans l'amélioration des performances des modèles. Cette étape a permis de sélectionner les meilleures configurations pour maximiser la précision et garantir la robustesse des résultats.

Enfin, ce projet a mis en lumière l'importance de l'évaluation rigoureuse des modèles à l'aide de métriques telles que la précision, le rappel et le F1-score, afin de s'assurer que les modèles soient adaptés aux exigences du projet. L'ensemble de ce travail fournit une base solide pour de futures recherches et développements, notamment l'intégration de techniques plus avancées telles que les modèles de réseaux de neurones et les architectures basées sur les transformateurs pour des performances encore plus élevées.

Cette étude a démontré que l'analyse des avis de produits par des méthodes de Text Mining offre un potentiel important pour comprendre les attentes et les sentiments des consommateurs, permettant ainsi aux entreprises de mieux orienter leurs stratégies de produit et de marketing.